

# The RNA Newton Polytope and Learnability of Energy Parameters

Elmirasadat Forouzmand  
 Department of Computer Science  
 Wayne State University  
 Detroit, MI 48202  
 elmira@wayne.edu

Hamidreza Chitsaz\*  
 Department of Computer Science  
 Wayne State University  
 Detroit, MI 48202  
 chitsaz@wayne.edu

January 9, 2013

## Abstract

**Motivation:** Computational RNA structure prediction is a mature important problem which has received a new wave of attention with the discovery of regulatory non-coding RNAs and the advent of high-throughput transcriptome sequencing. Despite nearly two scores of research on RNA secondary structure and RNA-RNA interaction prediction, the accuracy of the state-of-the-art algorithms are still far from satisfactory. So far, researchers have proposed increasingly complex energy models and improved parameter estimation methods, experimental and/or computational, in anticipation of endowing their methods with enough power to solve the problem. The output has disappointingly been only modest improvements, not matching the expectations. Even recent massively featured machine learning approaches were not able to break the barrier.

**Approach:** The first step towards high accuracy structure prediction is to pick an energy model that is inherently capable of predicting each and every one of known structures to date. In this paper, we introduce the notion of *learnability* of the parameters of an energy model as a measure of such an inherent capability. We say that the parameters of an energy model are *learnable* iff there exists at least one set of such parameters that renders *every* known RNA structure to date the minimum free energy structure. We derive a necessary condition for the learnability and give a dynamic programming algorithm to assess it. Our algorithm computes the convex hull of the feature vectors of all feasible structures in the ensemble of a given input sequence. Interestingly, that convex hull coincides with the *Newton polytope* of the partition function as a polynomial in energy parameters. To the best of our knowledge, this is the first approach towards computing the RNA Newton polytope and a systematic assessment of the inherent capabilities of an energy model.

**Results:** We demonstrated the application of our theory to a simple energy model consisting of a weighted count of A-U and C-G base pairs. Our results show that this simple energy model satisfies the necessary condition for less than one third of the input unpseudoknotted sequence-structure pairs chosen from the RNA STRAND v2.0 database. For another one third, the necessary condition is barely violated, which suggests that augmenting this simple energy model with more features such as the Turner loops may solve the problem. The necessary condition is severely violated for 8%, which provides a small set of hard cases that require further investigation.

---

\*to whom correspondence should be addressed

# 1 Introduction

Computational RNA structure and RNA-RNA interaction prediction have always been important problems, particularly now that RNA has been shown to have key regulatory roles in the cell [1, 2, 3, 4, 5, 6, 7]. Furthermore, with the advent of synthetic biology at the whole organism level [8], high-throughput accurate RNA engineering algorithms are required for both *in vivo* and *in vitro* applications [9, 10, 11, 12, 13]. Since the dawn of RNA secondary structure prediction nearly two scores ago [14], the research community has proposed increasingly complex models and algorithms, hoping that refined features together with better methods to estimate their parameters would solve the problem. Early approaches considered mere base pair counting, followed by the Turner thermodynamics model which was a significant leap forward. Recently, massively feature-rich models empowered by parameter estimation algorithms have been proposed, but they provide only modest improvements.

Despite significant progress in the last three decades, made possible by the work of Turner and others [15] on measuring RNA thermodynamic energy parameters and the work of several groups on novel algorithms [16, 17, 18, 19, 20, 21, 22, 23, 24] and machine learning approaches [25, 26, 27], the RNA structure prediction accuracy has not reached a satisfactory level yet. Why is it so? Up to now, human intuition and computational convenience have lead the way. We believe that human intuition has to be equipped with systematic methods to assess the suitability of a given energy model. Surprisingly, there is not a single method to assess whether the parameters of an energy model are *learnable*. We say that the parameters of an energy model are *learnable* iff there exists at least one set of such parameters that renders *every* known RNA structure to date, determined through X-ray or NMR, the minimum free energy structure. Equivalently, we say that the parameters of an energy model are learnable iff 100% structure prediction accuracy can be achieved when the training and test sets are identical. The first step towards high accuracy structure prediction is to make sure that the energy model is inherently capable, i.e. its parameters are learnable. In this work, we provide a necessary condition for the learnability and an algorithm to verify it. To the best of our knowledge, this is the first approach towards a systematic assessment of the suitability of an energy model. Note that a successful RNA folding algorithm needs to have the generalization power to predict unseen structures as well. We do not deal with the generalization power in this work and leave it for future work.

## 2 Background

### 2.1 RNA Secondary Structure Models

An RNA secondary structure model is often a context free grammar together with a scoring function for either the rules, in the case of stochastic context free grammars (SCFG) [28], or the alphabet, in the case of thermodynamics models [15]. Such scoring functions induce scoring on the entire generated language. The word with optimal score then yields a predicted structure for the given sequence. For the sake of brevity, we focus on thermodynamics models in this paper, but it is obvious that our methods apply to other models including SCFG as well. In our context, the scoring function is the thermodynamics free energy. A secondary structure  $y$  of a nucleic acid is decomposed into loops; a free energy is associated with every loop in  $y$ ; and the total free energy  $G$  for  $y$  is the sum of loop free energies [15]. The same loop decomposition principle applies to interacting nucleic acids such that the total free energy  $G$  is still the sum of the free energies of loops and interaction components [22].

## 2.2 Estimation of Energy Parameters

Existing machine learning algorithms for parameter estimation in RNA structure prediction can be grouped into two categories:

- Likelihood-based methods, where the maximum likelihood (ML) principle is used to estimate the parameters of a probabilistic model, e.g. [25], and
- Large-margin methods, where the model parameters are estimated to maximize the margin between the score of the true structure and the second best structure. This has been done using an online passive-aggressive training algorithm [27] and Iterative Constraint Generation (CG) [29].

The likelihood-based techniques estimate the best *Gibbs* distribution, which not only assists in predicting the best secondary structure but also is utilized in determining the thermodynamic parameters. The most successful method for learning the thermodynamics of RNA has been the maximum likelihood method, as in CONTRAfold [25], which maximizes the probability of RNA structures  $y$  given RNA sequences  $x$  for the training set  $D$ . That is, the conditional log likelihood of the training data (using the Boltzmann distribution) is maximized to estimate the best model parameters  $\mathbf{h}^* \in \mathbb{R}^k$ :

$$\mathbf{h}^* := \arg \max_{\mathbf{h}} L(D; \mathbf{h}) = \max_{\mathbf{h}} \sum_{(x,y) \in D} \log p(y|x, \mathbf{h}), \quad (1)$$

$$p(y|x, \mathbf{h}) := \frac{e^{-G(x,y,\mathbf{h})/RT}}{Q(x, \mathbf{h})}, \quad (2)$$

where  $k$  denotes the number of different motifs defined in the energy model,  $R$  is the gas constant,  $T$  is the absolute temperature,  $G(x, y, \mathbf{h})$  is the free energy, and

$$Q(x, \mathbf{h}) := \sum_{s \in \mathcal{E}(x)} e^{-G(x,s,\mathbf{h})/RT} \quad (3)$$

is the *partition function* [22, 20, 21] with  $\mathcal{E}(x)$  being the ensemble of possible structures of  $x$ . The free energy

$$G(x, s, \mathbf{h}) := \langle c(x, s), \mathbf{h} \rangle \quad (4)$$

is a linear function of the parameters  $\mathbf{h}$  where  $c(x, s) \in \mathbb{Z}^k$  is the features vector.

## 3 Learnability

The question that we ask before parameter estimation is: does there ever exist parameters  $\mathbf{h}^\dagger$  such that for every  $(x, y) \in D$ ,  $y = \arg \min_s G(x, s, \mathbf{h}^\dagger)$ ? If the answer to this question is no, then there is no hope that one can ever achieve 100% accuracy using the given model. The answer reveals inherent limitations of the model, which can be used to design improved models. We provide a necessary condition for the existence of  $\mathbf{h}^\dagger$  and a dynamic programming algorithm to verify it through computing the Newton polytope for every  $x$  in  $D$ . We will define the RNA Newton polytope below. Not only our algorithm provides a binary answer, it also quantifies the distance from the boundary.

## 4 Methods

### 4.1 Necessary Condition for Learnability

Let  $(x, y) \in D$  and  $\mathbf{h}^\dagger \in \mathbb{R}^k$ . Assume  $y$  minimizes  $G(x, s, \mathbf{h}^\dagger)$  as a function of  $s$ . In that case

$$G(x, y, \mathbf{h}^\dagger) \leq G(x, s, \mathbf{h}^\dagger), \quad \forall s \in \mathcal{E}(x). \quad (5)$$

Replacing (4) above,

$$\langle c(x, y), \mathbf{h}^\dagger \rangle \leq \langle c(x, s), \mathbf{h}^\dagger \rangle, \quad \forall s \in \mathcal{E}(x) \quad (6)$$

$$0 \leq \langle c(x, s) - c(x, y), \mathbf{h}^\dagger \rangle, \quad \forall s \in \mathcal{E}(x). \quad (7)$$

Define the *feature ensemble* of sequence  $x$  by

$$\mathcal{F}(x) := \{c(x, s) \mid s \in \mathcal{E}(x)\} \subset \mathbb{Z}^k. \quad (8)$$

In that case, (7) implies that

$$0 \leq \langle \mathcal{F}(x) - c(x, y), \mathbf{h}^\dagger \rangle. \quad (9)$$

We call the convex hull of  $\mathcal{F}(x)$  the *Newton polytope* of  $x$ ,

$$\mathcal{N}(x) := \text{conv} \{ \mathcal{F}(x) \} \subset \mathbb{R}^k. \quad (10)$$

We remind the reader that the convex hull of a set, denoted by ‘conv’ hereby, is the minimal convex set that fully contains the set. The reason for naming this polytope the Newton polytope will be made clear below. Inequality (9) implies that  $c(x, y) \in \partial \mathcal{N}(x)$  is on the boundary of the convex hull of the feature ensemble of  $x$  with a support hyperplane normal to  $\mathbf{h}^\dagger$ . Therefore, we have the following theorem.

**Theorem 1.** *Let  $(x, y) \in D$  and  $0 \neq \mathbf{h}^\dagger \in \mathbb{R}^k$ . Assume  $y$  minimizes  $G(x, s, \mathbf{h}^\dagger)$  as a function of  $s$ . In that case,  $c(x, y) \in \partial \mathcal{N}(x)$ , i.e. the feature vector of  $(x, y)$  is on the boundary of the Newton polytope of  $x$ .*

*Proof.* To the contrary, suppose  $c(x, y)$  is in the interior of  $\mathcal{N}(x)$ . Therefore, there is an open ball of radius  $\delta > 0$  centered at  $c(x, y)$  completely contained in  $\mathcal{N}(x)$ , i.e.

$$B_\delta(c(x, y)) \subset \mathcal{N}(x). \quad (11)$$

Let

$$p = c(x, y) - (\delta/2) \frac{\mathbf{h}^\dagger}{\|\mathbf{h}^\dagger\|}.$$

It is clear that  $p \in B_\delta(c(x, y)) \subset \mathcal{N}(x)$  since  $\|p - c(x, y)\| = \delta/2 < \delta$ . Therefore,  $p$  can be written as a convex linear combination of the feature vectors in  $\mathcal{F}(x) = \{v_1, \dots, v_N\}$ , i.e.

$$\exists \alpha_1, \dots, \alpha_N \geq 0 : \quad \alpha_1 v_1 + \dots + \alpha_N v_N = p, \quad (12)$$

$$\alpha_1 + \dots + \alpha_N = 1. \quad (13)$$

Note that

$$\langle p - c(x, y), \mathbf{h}^\dagger \rangle = -(\delta/2) \|\mathbf{h}^\dagger\| < 0. \quad (14)$$

Therefore, there is  $1 \leq i \leq N$ , such that  $\langle v_i - c(x, y), \mathbf{h}^\dagger \rangle < 0$  for otherwise,

$$\langle p - c(x, y), \mathbf{h}^\dagger \rangle = \sum_{i=1}^N \alpha_i \langle v_i - c(x, y), \mathbf{h}^\dagger \rangle \geq 0, \quad (15)$$

which would be a contradiction with (14). It is now sufficient to note that  $v_i \in \mathcal{F}(x)$  and  $\langle v_i - c(x, y), \mathbf{h}^\dagger \rangle < 0$  which is a contradiction with (9).  $\square$   $\square$

**Corollary 1** (Necessary Condition for the Learnability). *For  $(x, y) \in D$ , a necessary but not sufficient condition for the existence of  $\mathbf{h}^\dagger$  such that  $y$  minimizes  $G(x, s, \mathbf{h}^\dagger)$  as a function of  $s$  is that  $c(x, y)$  lies on the boundary of  $\mathcal{N}(x)$  the Newton polytope of  $x$ .*

## 4.2 Relation to the Newton Polytope

In addition to  $D$  the set of experimentally determined structures, we often have a repository of thermodynamic measurements, e.g. melting curves, that can help better estimate the energy parameters. Such measurements often relate to the energy parameters through equations involving the partition function and its derivatives with respect to temperature [22]. We show that with a change of variables, the partition function becomes a polynomial. Therefore, such equations become a system of polynomial equations the solving of which algebraically requires computation of the Newton polytope of each polynomial [30, 31]. Recall the partition function defined in (3) and energy in (4), and conclude

$$Q(x, \mathbf{h}) = \sum_{s \in \mathcal{E}(x)} e^{-\langle c(x, s), \mathbf{h} \rangle / RT}. \quad (16)$$

Let  $c(x, s) = (c_1(x, s), \dots, c_k(x, s))$  and  $\mathbf{h} = (\mathbf{h}_1, \dots, \mathbf{h}_k)$ . Define new variables

$$Z_i := e^{-\mathbf{h}_i / RT}, \quad 1 \leq i \leq k, \quad (17)$$

and replace them in (16). We obtain the partition function

$$Q(x, Z) = \sum_{s \in \mathcal{E}(x)} Z^{c(x, s)}, \quad (18)$$

in the form of a polynomial in  $\mathbb{R}[Z]$  where

$$Z^{c(x, s)} := \prod_{i=1}^k Z_i^{c_i(x, s)} \quad (19)$$

is a monomial as  $0 \leq c_i(x, s) \in \mathbb{Z}$ . The Newton polytope of  $Q$  is defined to be the convex hull of the monomials power vectors, i.e.

$$\text{Newton}\{Q(x, Z)\} := \text{conv}(\{c(x, s) \mid s \in \mathcal{E}(x)\}) = \mathcal{N}(x). \quad (20)$$

That is why we call  $\mathcal{N}(x)$  the Newton polytope of  $x$ .

### 4.3 RNA Newton Polytope Algorithm

We give a dynamic programming algorithm to compute the Newton polytope for a given nucleic acid sequence  $x$ . Denote the length of  $x$  by  $L$  and the  $i^{th}$  nucleotide in  $x$  by  $n_i$ . Denote the subsequence of  $x$  from the  $i^{th}$  to the  $j^{th}$  nucleotide, inclusive of ends, by  $n_i \cdots n_j$ . The following lemma allows us to formulate a divide-and-conquer strategy for computing the Newton polytope, which will in turn lead to our dynamic programming algorithm.

**Lemma 1.** *Let  $f$  and  $g$  be two polynomials in  $\mathbb{R}[Z]$ . The Newton polytope of the product of  $f$  and  $g$  is the Minkowski sum of individual Newton polytopes, and the Newton polytope of the sum of  $f$  and  $g$  is the convex hull of the union of individual Newton polytopes, i.e.*

$$Newton(fg) = Newton(f) \oplus Newton(g), \quad (21)$$

$$Newton(f + g) = \text{conv}\{Newton(f) \cup Newton(g)\}, \quad (22)$$

in which  $\oplus$  represents the Minkowski sum of two polytopes [30].

This lemma allows us to use the same divide-and-conquer strategy that was used for calculating the partition function [22, 20, 21]. We can use the same recursions (grammar) as in the partition function algorithm but with the Minkowski sum  $\oplus$  instead of multiplication, convex hull of union instead of summation, and the corresponding feature vector  $c$  instead of  $e^{-\langle c, \mathbf{h} \rangle / RT}$ . Furthermore, since union is invariant with respect to repetition of points, the dynamic programming is allowed to be redundant, or equivalently the grammar is allowed to be ambiguous. Hence, any complete RNA structure or RNA-RNA interaction prediction dynamic programming algorithm can be transformed into a Newton polytope algorithm by replacing the energy with the corresponding feature vector, summation with the Minkowski sum  $\oplus$ , and minimization with the convex hull of union.

**As explained above, we transform any complete partition function or structure prediction dynamic programming algorithm, for single RNA, RNA-RNA interaction, or multiple interacting RNAs, into a Newton polytope algorithm. For the sake of illustration, we explicitly spell below only the case of single RNA with separate A-U and C-G base pair counting energy model. All the other cases are quite trivially obtained following the transformations above.**

In this case, the feature vector  $c(x, s) = (c_1(x, s), c_2(x, s))$  is two dimensional:  $c_1(x, s)$  is the number of A-U, and  $c_2(x, s)$  the number of C-G base pairs in  $s$ . Our dynamic programming algorithm starts by computing the Newton polytope for all unit length subsequences, followed by all length two subsequences, ..., up to the Newton polytope for the entire sequence  $x$ . We denote the Newton polytope of the subsequence  $n_i \cdots n_j$  by  $\mathcal{N}(i, j)$ , i.e.

$$\mathcal{N}(i, j) := \mathcal{N}(n_i \cdots n_j). \quad (23)$$

The following dynamic programming will yield the result

$$\mathcal{N}(i, j) = \text{conv} \left[ \bigcup \left\{ \begin{array}{ll} \mathcal{N}(i, \ell) \oplus \mathcal{N}(\ell + 1, j), & i \leq \ell \leq j - 1 \\ \{(1, 0)\} \oplus \mathcal{N}(i + 1, j - 1) & \text{if } n_i n_j = \text{AU|UA} \\ \{(0, 1)\} \oplus \mathcal{N}(i + 1, j - 1) & \text{if } n_i n_j = \text{CG|GC} \end{array} \right. \right], \quad (24)$$

with the base case  $\mathcal{N}(i, i) = \{(0, 0)\}$ .

There are two different approaches for polytope representation: (i) vertex representation, which is a set of points, and (ii) half plane representation, which is a set of linear inequalities. The former is often called  $\mathcal{V}$ -representation and the latter  $\mathcal{H}$ -representation. Although they are equivalent,

and there are algorithms to transform one into the other, computing Minkowski sum is more convenient with the  $\mathcal{V}$ -representation, and convex hull of union works more efficiently with the  $\mathcal{H}$ -representation. The choice of representation and algorithms will affect the running time. In this paper, we use the  $\mathcal{V}$ -representation.

#### 4.4 Verification of the Necessary Condition

Upon computation of  $\mathcal{N}(x)$  and  $c(x, y)$ , the feature vector of the experimentally determined structure, it remains to verify whether  $c(x, y) \in \partial\mathcal{N}(x)$ . Often,  $\mathcal{N}(x)$  is represented by its vertices ( $\mathcal{V}$ -representation) or its confining half planes ( $\mathcal{H}$ -representation), two equivalent representations that can be transformed into one another. In an  $\mathcal{H}$ -representation,  $c(x, y)$  is on the boundary of  $\mathcal{N}(x)$  iff there is at least one confining plane on which  $c(x, y)$  lies. This is true because  $c(x, y) \in \mathcal{N}(x)$  anyways. Therefore, the necessary condition can be easily checked by checking membership of  $c(x, y)$  in every confining plane. Since the vertices of  $\mathcal{N}(x)$  are on the integer lattice, all calculations are rational and hence can be performed exactly.

#### 4.5 Dataset

We used 1720 unpsuedoknotted RNA sequence-structure pairs from RNA STRAND v2.0 database as our dataset  $D$ . RNA STRAND v2.0 contains known RNA secondary structures of any type and organism, particularly with and without pseudoknots. To the best of our knowledge, RNA STRAND v2.0 is the most comprehensive collection of known RNA secondary structures to date [32]. There are 2334 pseudoknot-free RNAs in the RNA STRAND database. We sorted them based on their length and selected the first 1720 ones, whose lengths vary between 4 and 123 nt. We excluded pseudoknotted structures because our current implementation is incapable of considering pseudoknots. Some sequences in the dataset allow only A-U base pairs (not a single C-G pair), in which case the Newton polytope degenerates into a line.

#### 4.6 Implementation

We implemented the dynamic programming in (24) using MATLAB convex hull function which is based on the quickhull algorithm [33]. As mentioned above, we used the  $\mathcal{V}$ -representation and computed the Minkowski sum by direct pairwise summation of vertices. More precisely, for two convex polytopes  $P$  with vertices  $p_1, \dots, p_a$  and  $Q$  with vertices  $q_1, \dots, q_b$ , the vertices of  $P \oplus Q$  are  $p_i + q_j$  for  $1 \leq i \leq a$  and  $1 \leq j \leq b$ . To verify the necessary condition, i.e. whether the experimentally determined feature vector lies on the boundary of the Newton polytope, we calculated the distance of the feature vector from the boundary of the polytope using the ‘p\_poly\_dist’ MATLAB function [34]. A zero distance corresponds to the case where the feature vector lies on the boundary, i.e. the condition is satisfied, and a positive distance to the case where the feature vector is in the interior of the Newton polytope. We normalized the distance by square root of the area of the polytope, which is a planar polygon in this case. The normalized distance quantifies how far the feature vector is from the boundary. We parallelized our MATLAB code using MATLAB ‘parfor’. We ran experiments on the JSLQ nodes of the Wayne State Grid in however non-parallel mode due to lack of support. The length of input RNA sequences varied between 4 and about 120 nt. For the smallest ones, our program took a fraction of a second and for the longest ones it took less than 10 minutes to run on a 2.4GHz Dual Core AMD Opteron CPU with at most 16 GB RAM.



## 5 Results

For each strand of RNA, the distance between  $c(x, y)$ , the real feature vector of the secondary structure and the computed convex hull,  $\mathcal{N}(x)$ , is calculated using [34]. We denote this distance by  $r(x)$  here. The necessary condition for the learnability is satisfied if  $r(x) = 0$  for all  $x$  in the dataset, which shows that the observed feature vector lies on the boundary of  $\mathcal{N}(x)$ .

Fig. 1 illustrates the secondary structure of h.5.b.E.coli.hlxnum, an example in the dataset, and its Newton polygon. This RNA is 120 nt long, and its final Newton polygon has 12 vertices. The native feature vector for this RNA has  $r = 0$  which means that it lies on the boundary of the polygon. Fig. 2 illustrates another example, Hammerhead Ribozyme (Type I) whose length is 115 nt. Its Newton polygon has 15 vertices, and the calculated  $r$  is 5 for this case. The big hairpin loop in this case suggests that the dominant energetic features are not base pairs, and this is consistent with the longer distance observed between  $\mathcal{N}(x)$  and  $c(x, y)$ .

Out of 1720 RNA sequences in our experiment, for 490 (28%) the native feature vector is exactly on the boundary of the Newton polygon. For the remaining 1230 sequences, the native feature vector settles inside their Newton polygon. Fig. 3 demonstrates the histogram of  $r(x)$  for the input dataset. For 516 (30%) of input sequences,  $r$  is non-zero but less than 1, which suggests that the dominant energetic features in those cases are canonical base pairs. For 579 (34%) sequences,  $r$  is between 1 and 5. In the remaining 135 cases (8%),  $r$  goes larger than 5. The second plot in Fig. 3 shows the normalized distance histogram. The square root of polygon area is used as the normalization factor. For 712 (41%) strands, this number is no more than 0.05.

The relation between the number of vertices and the length of RNA strand is shown in the third plot of Fig. 3. The maximum number of vertices is 15 which happens for three different strands, each 115 nt long. The minimum number of vertices is two; however, it is clear that no polygon exists with two vertices. In those cases, the RNA admits only A-U base pairs or only C-G base pairs and not both of them. The resulting polygon is just a line, and  $c(x, y)$  always lies on the boundary of  $\mathcal{N}(x)$ .

## 6 Conclusion and Future Work

We introduced the notion of learnability of the parameters of an energy model as a measure of its inherent capability. We derived a necessary condition for the learnability and gave a dynamic programming algorithm to assess it. Our algorithm computes the convex hull of the feature vectors of all feasible structures in the ensemble of a given input sequence. Also, that convex hull coincides with the *Newton polytope* of the partition function as a polynomial in transformed energy parameters.

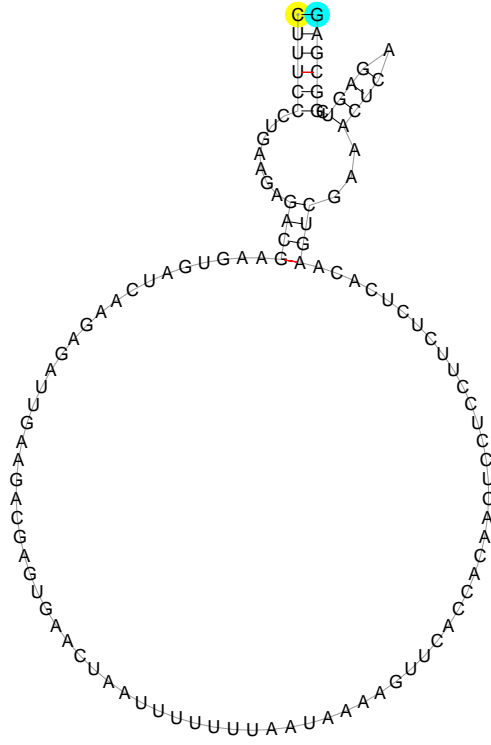
Our theory applied to a simple energy model that counts A-U and C-G base pairs separately revealed that about one third of chosen known structures could potentially be predicted using this simple energy model. For another one third, the necessary condition is barely violated, which suggests that augmenting this energy model with more features is expected to solve the problem. The condition is severely violated for 8% of sequences, which will be the subject of future investigation. The twilight zone is also interesting and requires deeper examination.

The Newton polytope lies in the core of computer algebra for solving polynomial equations. Therefore, we envision applications of our RNA Newton polytope in symbolic estimation of energy parameters. Sufficient conditions for the learnability, and also assessing the generalization power of an energy model remain for future work.

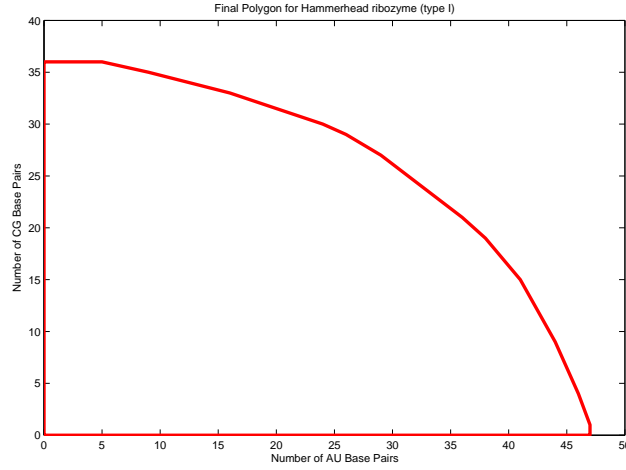




9



(a) Secondary structure of Hammerhead Ribozyme (Type I).



(b) The Newton polygon of Hammerhead Ribozyme (Type I), with 15 vertices.

Figure 2: The secondary structure and Newton polygon of Hammerhead Ribozyme (Type I) in RNA STRAND v2.0 [32]. The native feature vector does not lie on the boundary in this case.

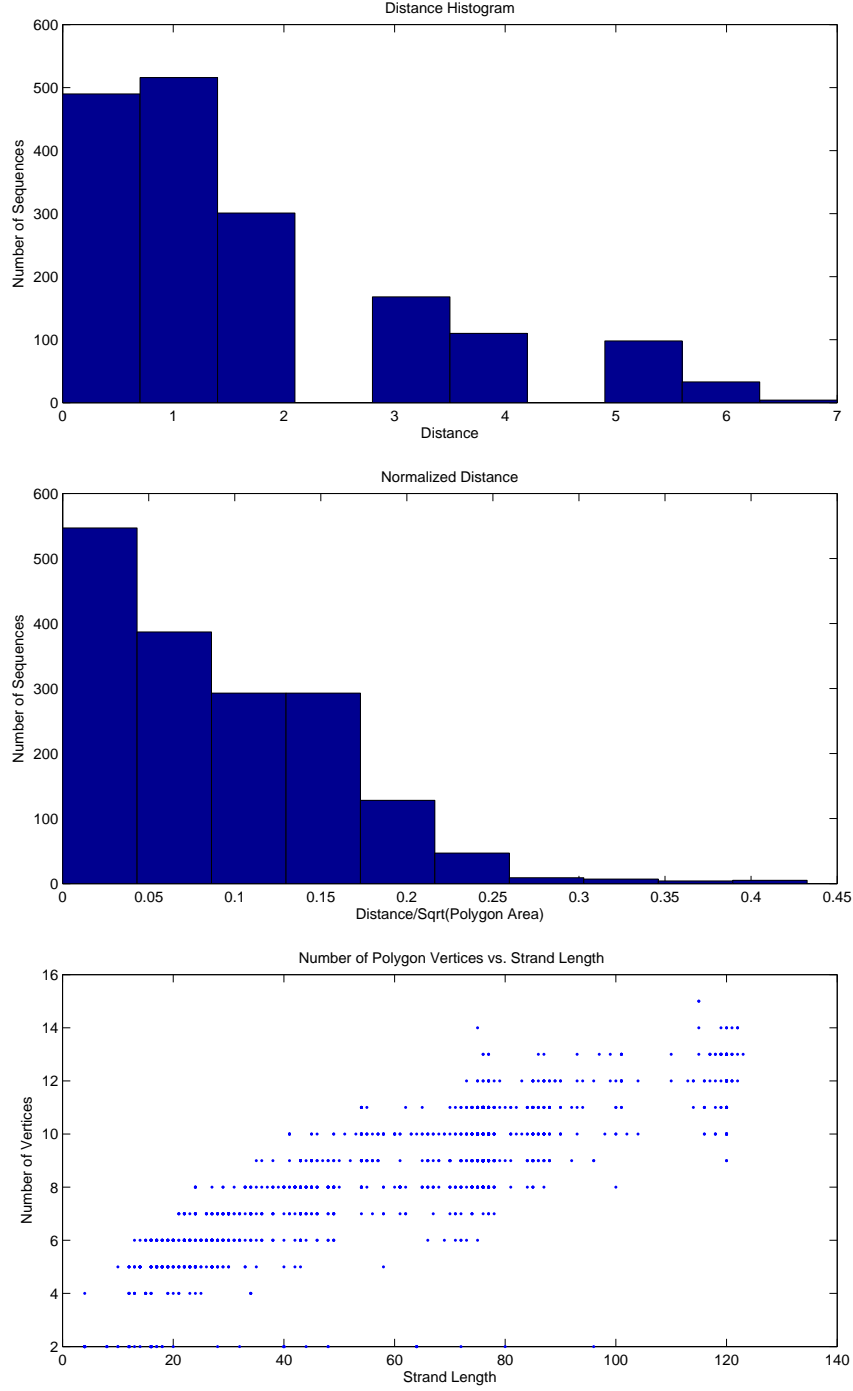


Figure 3: (Top) Histogram of  $r$ . (Middle) Histogram of  $r(x)/\sqrt{\text{area}(\mathcal{N}(x))}$ . (Bottom) Scatter plot of the number of vertices of the Newton polygon as a function of sequence length.

## References

- [1] Gisela Storz. An expanding universe of noncoding RNAs. *Science*, 296(5571):1260–3, 2002.

- [2] David P. Bartel. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, 116(2):281–97, 2004.
- [3] Gregory J. Hannon. RNA interference. *Nature*, 418(6894):244–51, 2002.
- [4] Phillip D. Zamore and Benjamin Haley. Ribo-gnome: the big world of small RNAs. *Science*, 309(5740):1519–24, 2005.
- [5] E.G. Wagner and K. Flardh. Antisense RNAs everywhere? *Trends Genet.*, 18:223–226, May 2002.
- [6] S. Brantl. Antisense-RNA regulation and RNA interference. *Bioch. Biophys. Acta*, 1575(1-3):15–25, 2002.
- [7] Susan Gottesman. Micros for microbes: non-coding regulatory RNAs in bacteria. *Trends in Genetics*, 21(7):399–404, 2005.
- [8] Daniel G. Gibson, John I. Glass, Carole Lartigue, Vladimir N. Noskov, Ray-Yuan Chuang, Mikkel A. Algire, Gwynedd A. Benders, Michael G. Montague, Li Ma, Monzia M. Moodie, Chuck Merryman, Sanjay Vashee, Radha Krishnakumar, Nacyra Assad-Garcia, Cynthia Andrews-Pfannkoch, Evgeniya A. Denisova, Lei Young, Zhi-Qing Qi, Thomas H. Segall-Shapiro, Christopher H. Calvey, Prashanth P. Parmar, Clyde A. Hutchison, Hamilton O. Smith, and J. Craig Venter. Creation of a bacterial cell controlled by a chemically synthesized genome. *Science*, 329(5987):52–56, 2010.
- [9] N.C. Seeman. From genes to machines: DNA nanomechanical devices. *Trends Biochem. Sci.*, 30:119–125, Mar 2005.
- [10] N. C. Seeman and P. S. Lukeman. Nucleic acid nanostructures: bottom-up control of geometry on the nanoscale. *Reports on Progress in Physics*, 68:237–270, January 2005.
- [11] F.C. Simmel and W.U. Dittmer. DNA nanodevices. *Small*, 1:284–299, Mar 2005.
- [12] S. Venkataraman, R.M. Dirks, P.W. Rothmund, E. Winfree, and N.A. Pierce. An autonomous polymerization motor powered by DNA hybridization. *Nat Nanotechnol*, 2:490–494, Aug 2007.
- [13] P. Yin, R.F. Hariadi, S. Sahu, H.M. Choi, S.H. Park, T.H. Labean, and J.H. Reif. Programming DNA tube circumferences. *Science*, 321:824–826, Aug 2008.
- [14] P. N. Borer, B. Dengler, I. Tinoco, and O. C. Uhlenbeck. Stability of ribonucleic acid double-stranded helices. *J. Mol. Biol.*, 86(4):843–853, Jul 1974.
- [15] D.H. Mathews, J. Sabina, M. Zuker, and D.H. Turner. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.*, 288:911–940, May 1999.
- [16] R. Nussinov, G. Piecznik, J. R. Grigg, and D. J. Kleitman. Algorithms for loop matchings. *SIAM Journal on Applied Mathematics*, 35:68–82, 1978.
- [17] M. S. Waterman and T. F. Smith. RNA secondary structure: A complete mathematical analysis. *Math. Biosci.*, 42:257–266, 1978.
- [18] Michael Zuker and Patrick Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Research*, 9(1):133–148, 1981.

- [19] E. Rivas and S.R. Eddy. A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J. Mol. Biol.*, 285:2053–2068, Feb 1999.
- [20] Robert M. Dirks and Niles A. Pierce. A partition function algorithm for nucleic acid secondary structure including pseudoknots. *Journal of Computational Chemistry*, 24(13):1664–1677, 2003.
- [21] J.S. McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29:1105–1119, 1990.
- [22] Hamidreza Chitsaz, Raheleh Salari, S.Cenk Sahinalp, and Rolf Backofen. A partition function algorithm for interacting nucleic acid strands. *Bioinformatics*, 25(12):i365–i373, 2009.
- [23] Hamidreza Chitsaz, Rolf Backofen, and S.Cenk Sahinalp. biRNA: Fast RNA-RNA binding sites prediction. In *Workshop on Algorithms in Bioinformatics (WABI)*, volume 5724 of *LNBI*. Springer, 2009.
- [24] S.H. Bernhart, H. Tafer, U. Mückstein, C. Flamm, P.F. Stadler, and I.L. Hofacker. Partition function and base pairing probabilities of RNA heterodimers. *Algorithms Mol Biol*, 1:3, 2006.
- [25] C. B. Do, D. A. Woods, and S. Batzoglou. CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics*, 22:90–98, Jul 2006.
- [26] M. Andronescu, A. Condon, H. H. Hoos, D. H. Mathews, and K. P. Murphy. Computational approaches for RNA energy parameter estimation. *RNA*, 16:2304–2318, Dec 2010.
- [27] Shay Zakov, Yoav Goldberg, Michael Elhadad, and Michal Ziv-Ukelson. Rich parameterization improves RNA structure prediction. In Vineet Bafna and S. Sahinalp, editors, *Proceedings of the 15th Annual international conference on Research in Computational Molecular Biology*, volume 6577 of *Lecture Notes in Computer Science*, pages 546–562. Springer Berlin-Heidelberg, 2011.
- [28] S. R. Eddy and R. Durbin. RNA sequence analysis using covariance models. *Nucleic Acids Res.*, 22(11):2079–2088, Jun 1994.
- [29] M. Andronescu, A. Condon, H. H. Hoos, D. H. Mathews, and K. P. Murphy. Efficient parameter estimation for RNA secondary structure prediction. *Bioinformatics*, 23:19–28, Jul 2007.
- [30] I. Emiris. *Sparse Elimination and Applications in Kinematics*. PhD thesis, UC Berkeley, Berkeley, CA, 1994.
- [31] Ioannis Z. Emiris and John F. Canny. Efficient incremental algorithms for the sparse resultant and the mixed volume. *J. Symbolic Computation*, 20:14–9, 1995.
- [32] M. Andronescu, V. Bereg, H. H. Hoos, and A. Condon. RNA STRAND: the RNA secondary structure and statistical analysis database. *BMC Bioinformatics*, 9:340, 2008.
- [33] C. Bradford Barber, David P. Dobkin, and Hannu Huhdanpaa. The quickhull algorithm for convex hulls. *ACM Trans. Math. Softw.*, 22(4):469–483, December 1996.
- [34] Michael Yoshpe. Distance from a point to a 2D polygon. <http://www.mathworks.com/matlabcentral/fileexchange/12744-distance-from-a-point-to-polygon> 2006. [Online].